

BGN/PCGN NATIONAL ROMANIZATION SYSTEM FOR AFGHANISTAN

BGN/PCGN 2007 System

This romanization system agreed by BGN and PCGN in November 2007, accommodates the linguistic complexity of Afghanistan as manifest in its geographical names. The following tabulation shows the original Perso-Arabic script with accompanying Unicode value (columns 1a and b), the Yaghoubi romanization¹ (column 2), the BGN/PCGN romanization with accompanying Unicode value (columns 3a and b), an English phonetic example (column 4), and an example toponym (columns 5b and c).

The tables function as both a romanization system for Afghanistan (i.e. with access to the original script, these tables can be applied to get a standardized Roman result - moving from columns 1 to 3) and as a means of converting the available Yaghoubi Roman-script spellings, as appear on the Fairchild Aerial Surveys map series, to standard BGN/PCGN spellings (moving from columns 2 to 3).

The points used in Arabic to mark short vowels and certain other diacritical marks are infrequently written in Afghanistan. Consequently, a reference source may sometimes be required to aid correct identification of the standard spellings and proper vowels and elimination of dialectal and idiosyncratic variations. In the interests of clarity, the example columns show script with vowel pointing from Arabic to indicate the short vowels that are included alongside the unpointed form that will usually be encountered. However it should be noted that the pronunciation of short vowels will vary.

(Note: it is recommended that a font such as Scheherazade, available from www.sil.org, which includes the Unicode extended Arabic sub-range, be used to view this system².)

¹The Yaghoubi romanization system was developed in 1959 by Muzaffarud Din Yaqubi (commonly seen as Yaghoubi). It is a native official system designed to reflect Afghan names, both Dari and Pashto, and both pronunciation and genuine linguistic truth.

²Please note that the identification of a particular font does not represent an endorsement of any specific product or manufacturer.

Table 1: Consonant Characters

	Script				Unicode codepoint (Independent)	Yaghoubi	Romanization	BGN/PCGN Unicode codepoint (lower case)	English phonetic	Example		
	Final	Medial	Initial	Independent						Unpointed script	Pointed script	Roman script
1	ا		ا		0627		<i>See note 1</i>	-		<i>See note 1</i>		
2	ب	ب	ب	ب	0628	b	b	0062	\b\ as in <i>bob</i>	بغلان	بَغْلَان	Baghlān
3	پ	پ	پ	پ	067E	p	p	0070	\p\ as in <i>pep</i>	پوتکی	پوْتَكِي	Pōṭakay
4	ت	ت	ت	ت	062A	t̄	t	0074	\t\ as in <i>tall</i>	شیرین تگاب	شیرین تَغَاب	Shīrīn Tagāb
5	<i>characters not available in Unicode³</i>			ت	067C	t	ṭ	1E6F	retroflexive \ṭ\	کوت	کوت	Kōṭ
6	ث	ث	ث	ث	062B	ṣ	ṣ	0073+0304	\s\ as in <i>sand</i>	ثابر	ثَابِر	Ṣābir
7	ج	ج	ج	ج	062C	j	j	006A	\j\ as in <i>jazz</i>	جلال آباد	جَلَال آباد	Jalālābād
8	چ	چ	چ	چ	0686	č	ch	0063/0068	\ch\ as in <i>chat</i>	چاریکار	چاریکار	Chārikār
9	ح	ح	ح	ح ⁴	0681	j̄	dz ^{Note2}	0064+007A	\dz\ as in <i>adze</i>	خدران	خُدْرَان	Dzadrān
10	ح	ح	ح	خ	0685	c	ts ^{Note2}	0074+0073	\ts\ as in <i>gnats</i>	خوکی	خَوَكِي	Tsowkēy
11	ح	ح	ح	ح	062D	ħ	ħ	1E29	\h\ as in <i>hot</i>	حضرت امام	حَضْرَتِ اِمَام	Ḥazrat-e Imām
12	خ	خ	خ	خ	062E	kh	kh	006B+0068	\ch\ as in German <i>ach</i>	خوست	خوست	Khōst
13	د		د		062F	ḍ	d	0064	\d\ as in <i>dog</i>	سپین بولدک	سپین بَوْلَدَك	Spīn Bōldak
14	<i>character not available in Unicode³</i>			د	0689	d	ḍ	1E0F	retroflexive \ḍ\	دند و پتان	دَنْد و پَتَان	Ḍaṇḍ wa Patān
15	ذ		ذ		0630	z̄	z̄	007A+0304	\z\ as in <i>zero</i>	گذرگاه نور	گُذْرَگَاه نُوْر	Gužargāh-e Nūr
16	ر		ر		0631	r̄	r	0072	\r\ as in <i>run</i>	کندهار	کَنْدِهَار	Kandahār
17	<i>character not available in Unicode³</i>			ر	0693	r	ṛ	1E5F	retroflexive \ṛ\	اندر	اَنْدِر	Andar
18	ز		ز		0632	z	z	007A	\z\ as in <i>zero</i>	کندز	کُنْدُز	Kunduz
19	ژ		ژ		0698	ž	zh	007A+0068	\zh\ as in <i>azure</i>	میر اسلم ژرنده	میر اَسْلَم ژَرَنْدَه	Mīr Aslam Zhrandah

³ These characters are not available with a single Unicode codepoint, so cannot be displayed here. When typing, the independent character's codepoint will automatically display with the appropriate word-medial or word-final form where so appearing in a word.

⁴ The variant form ح is seen infrequently, and does not have a single Unicode codepoint.

	Script				Unicode codepoint (Independent)	Yaghoubi	Romanization	BGN/PCGN Unicode codepoint (lower case)	English phonetic	Example		
	Final	Medial	Initial	Independent						Unpointed script	Pointed script	Roman script
20	character not available in Unicode ³			ز	0696	ž	zh	007A+035F+0068	retroflexive \zh\	بریره	بریره	Zhīrah
21	س	س	س	س	0633	s	s	0073	\s\ as in <i>sand</i>	سمنگان	سَمَنگان	Samangān
22	ش	ش	ش	ش	0634	š	sh	0073+0068	\sh\ as in <i>ship</i>	مزار شریف	مَزَارِ شَرِيف	Mazār-e Sharīf
23	characters not available in Unicode ³			ښ	069A	š	sh	0073+035F+0068	retroflexive \sh\	کښته کلا	کښته کلا	Kshêtah Kalā
24	ص	ص	ص	ص	0635	ṣ	ṣ	015F	\s\ as in <i>sand</i>	قیصار	قَيْصار	Qayṣār
25	ض	ض	ض	ض	0636	ḏ	z	1E95	\z\ as in <i>zero</i>	فیض آباد	فَيْض آباد	Faīzābād ^{Note 3}
26	ط	ط	ط	ط	0637	t̤	ṭ	0163	\t\ as in <i>tall</i>	حضرت سلطان	حَضْرَتِ سُلْطَان	Ḥazrat-e Sulṭān
27	ظ	ظ	ظ	ظ	0638	ẓ	ẓ	007A+0327	\z\ as in <i>zero</i>	ظاهر کلا	ظَاهِر کلا	Zāhir Kalā
28	ع	ع	ع	ع	0639	ʿ	ʿ	2018	no sound	پل علم	پُلِ عَلم	Pul-e ‘Alam
29	غ	غ	غ	غ	063A	gh	gh	0067+0068	no English equivalent like <i>r\</i> in French <i>mer</i>	غزني	غَزْنِي	Ghaznī
30	ف	ف	ف	ف	0641	f	f	0066	\f\ as in <i>fat</i>	مزار شریف	مَزَارِ شَرِيف	Mazār-e Sharīf
31	ق	ق	ق	ق	0642	q	q	0071	no English equivalent closest sound is guttural <i>\k\</i>	قیصار	قَيْصار	Qayṣār
32	ک	ک	ک	ک	06A9	k	k	006B	\k\ as in <i>kitten</i>	کندهار	کَنْدَهَار	Kandahār
33	گ	گ	گ	گ, ځ	06AF	g	g	0067	\g\ as in <i>gain</i>	گردېز	گَرْدېز	Gardēz
34	ل	ل	ل	ل	0644	l	l	006C	\l\ as in <i>land</i>	کابل	کَابُل	Kābul
35	م	م	م	م	0645	m	m	006D	\m\ as in <i>mint</i>	میمنه	مَيْمَنَه	Maīmanah
36	ن	ن	ن	ن	0646	n	n	006E	\n\ as in <i>night</i>	خان آباد	خان آباد	Khānābād
37	characters not available in Unicode ³			ښ	06BC	ɳ	ɳ	1E49	retroflexive \n\	مانی	مانی	Māṇēy
38	و		و	و	0648	w	w	0077	\w\ as in <i>wait</i>	واخان	واخان	Wākhān
39	ه	ه, ه	ه	ه	0647	h	h	0068	\h\ as in <i>hot</i>	کندهار	کَنْدَهَار	Kandahār
40	ی, ی	ی	ی	ی, ی	0649, 064A	y	y	0079	\y\ as in <i>yard</i>	ینگی قلعه	يَنْگِي قَلْعَه	Yangī Qal‘ah

Table 2: Vowel, diphthong and diacritical characters

	Script	Unicode value	Yaghoubi	BGN/PCGN	BGN/PCGN Unicode value	English Phonetic	Example		
							Unpointed script	Pointed script	Roman script
1	اَ	064E	a, â	a	0061	short \a\ as in <i>man</i>	جلال آباد	جَلال آباد	Jalālābād
2	اِ	0650	e, i	e, i ⁵	0065+0069	short \i\ or \e\ as in <i>sit</i> or <i>ten</i>	هرات پل حصار	هَرَات پُلِ حِصَار	Herāt, Pul-e Ḥiṣār
3	اُ	064F	o, u	o, u ⁶	006F, 0075	short \u\ or \o\ as in <i>put</i> or <i>no</i>	کابل	کَابُل	Kābul
4	اَو	0659	ə, ě	ê	00EA	Pashto vowel \ə\ as in <i>up</i> , <i>jug</i>	گردون	گَرْدُون	Gêrdōn
5	اِ	0627	ā	ā ^{see note 1}	0101	long \aa\ as in <i>car</i>	کابل	کَابِل	Kābul
6	آ	0622	ā	ā ^{see note 1}	0101	long \aa\ as in <i>car</i>	آب بند	آب بَند	Āb Band
7	و	0648	ō	ō	014D	long \o\ as in <i>boat</i> , <i>toe</i>	سپین بولدک	سپین بولَدک	Spīn Bōldak
8	و	0648	u, ū	ū	016B	long \oo\ as in <i>moon</i> , <i>rule</i>	بالا بلوک	بالا بُلُوک	Bālā Bulūk
9	او	0648	aw or āw	ow	006F+0077	diph. \ow\ as in <i>now</i> , <i>town</i>	جوزجان	جَوَزجان	Jowzjān
10	ی	06CC (064A)	i or ī	ī	012B	long \ee\ as in <i>meet</i> , <i>see</i>	غزنی، سپین	غَزْنی، سِپین	Ghaznī, Spīn
11	ی	06CC	ay or ē	ay or ē ⁸	0113, 0061+0079	diph. \ay\ as in <i>date</i> , <i>maid</i>	میوند، ریگ	مَیَوَند، رِیگ	Maywand, Rēg
12	ې	06D0	ē or ay	ē	0113	Pashto vowel as in <i>rain</i> , <i>tail</i>	گردېز	گَرْدېز	Gardēz

⁵ Both *e* and *i* are available to romanize this short vowel, depending on local usage and/or root language. In cases where the sound is uncertain, *i* is the default romanization in BGN/PCGN standardization procedures.

⁶ Both *o* and *u* are available to romanize this short vowel, depending on local usage and/or root language. In cases where the sound is uncertain, *u* is the default romanization in BGN/PCGN standardization procedures.

⁷ An *alif* with *maddah* (اِ) is written only in the initial position by BGN/PCGN standardization procedures, in keeping with Persian language family standards of use of the Arabic alphabet. The same letter written in a medial or final position is written ا (see note 1).

⁸ The character ی should be romanized *ay* or *ē* according to its root language or local pronunciation. In case of uncertainty a reference source (such as the Fairchild Aerial Surveys map series, or a BGN/PCGN approved policy document/list of recommended spellings) should be consulted.

	Script	Unicode value	Yaghoubi	BGN/PCGN	BGN/PCGN Unicode value	English Phonetic	Example		
							Unpointed script	Pointed script	Roman script
13	ی	06CC	ay or āy	ay or aī ⁹	0113, 0061+0079, 0061+012B	diph. \ɪ\ as in <i>my, eye</i>	میدان شهر	میدان شَهر	Māidān Shahr
14	ی	06CC	ā	á	00E1	long \aa\ as in <i>car</i>	دندِ سفلی	دَندِ سَفلی	Dand-e Suflá
15	ی	06CD	əy or ay	êy	00EA+0079	Pashto vowel \əy\ word ending	خوکی	خَوکی	Tsowkêy
16	ء	0621	'	'	2019	as in <i>day</i> or <i>year</i>	هوائی دگر	هَوائی دَگر	Hawā'ī Dagar
	ء as izāfah	0674	-i-, e, ī	-e, -ye	012B	as in <i>day</i> or <i>year</i> (glottal stop)	مزار شریف	مَزارِ شَرِیف	Mazār-e Sharīf
17	ئ	064A+0654	-	êy	00EA+0079	as in <i>shy</i>	- ¹⁰	-	-

Numerals									
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
0	1	2	3	4	5	6	7	8	9

Although Perso-Arabic script is written from right to left, numerical expressions, e.g. ١٩٦٨ → 1968, are written from left to right.

⁹ Both the combination *ay* and *aī* are available to romanize this character according to its root language or local pronunciation. In cases where the sound is uncertain *ay* is the default romanization in BGN/PCGN standardization procedures

¹⁰ Only occurs at the end of verbs, so unlikely to be encountered in geographical names.

NOTES

1. Alif (ا) should be romanized as follows:
 - a. Initially, it indicates that the word begins with a vowel or diphthong; the alif itself is not romanized, but rather the short vowel it “carries” is romanized; e.g., میرِ اَسْلَمِ ژرَنْدَه → Mīr Aslam Zhrandah
 - b. When it carries a maddah (آ) (see vowel table, row 6), it represents ā; e.g., اَب بَنْد → Āb Band.
 - c. Medially and finally it represents ā (see vowel table, row 5); e.g., مَانِي → Mānēy
 - d. Medially and finally in words of Arabic origin, alif may serve as the bearer of hamzah, e.g. رَأْس → ra’s.
2. Occasionally the letter sequences كه, زه, سه, and گه occur without intervening vowels. They may be romanized *k·h*, *z·h*, *s·h*, and *g·h* in order to differentiate these romanizations from the digraphs *kh*, *zh*, *sh*, and *gh*, which are used to represent the letters خ, ژ, ش, and غ. Additionally, the Pashto letters څ and ځ, routinely romanized *ts* and *dz*, may be alternatively romanized *t̄s* and *d̄z* when for special reasons it is desired that confusion be avoided with the character sequences تس (*ts*) and دز (*dz*), respectively.
3. The vagaries of written Afghan languages, as pertains to spacing and word division, are addressed as follows:

Spaces may be added to or subtracted from Afghan words written in Arabic script, for the purposes of standardization. This is particularly relevant when the words are hand-written, are rendered “artistically”, or express other such non-standard flourishes, as long as the sense of the toponym, word, or phrase is not compromised.

Romanized toponyms are typically divided into constituent words (spaces and other grammatical rules applied) when those words can stand independently, for purposes of standardization and minimization of confusion, particularly in situations where Afghan writers are inconsistent in their application of spacing and word breaks. When the Afghan word or suffix is only used in combination with other nouns or adjectives, then it should be appended to the preceding word in its romanization. This includes (but is not limited to) -*ābād*, -*zaī*, -*zādah*, -*tū*, -*wand*, -*gaī*, -*kaī*, -*pūr*, -*dēsh*, -*lar*, -*lī*, -*lū* and *ullāh*, as, for example, seen in *Raḥmatābād* (رحمت آباد) and *Raḥmatullāh* (رحمت الله), but *Raḥmat Khēl* (رحمتخیل) and *Raḥmat Shahr* (رحمتشهر).
4. The one-letter words د (Pashto) and و (Dari) are romanized *dê* and *wa*, respectively.
5. The word الله, meaning God, should always be romanized *Allāh*, except as specified in note 3. Note that the Unicode value FDF2 spells *Allāh*, but omits the *alif* in some common fonts, including Times New Roman. If in doubt, try in Arial Unicode MS to verify. Also note that the “dagger alif” (ا) above the second ل (lām) in the word الله, is not written but should be romanized *ā*, like a full-size alif.
6. In names of Arabic origin, the *l* of the definite article *al* is assimilated before the ‘sun letters’ *t*, *ṣ*, *d*, *z*, *r*, *z*, *s*, *sh*, *ṣ*, *z*, *ṭ*, *ṣ*, *l* and *n*. In its romanization, the article should be separated from the name it precedes and should not be capitalized except at the beginning of a name, e.g. جبل السراج → *Jabal as Sarāj*
7. In Arabic names, a *shaddah*, ّ is used to denote the doubling of a particular consonant character, e.g. مُحَمَّد → *Muḥammad*. However, in Pashto this ‘doubling’ is frequently omitted in both Perso-Arabic script and the resulting romanization. Guidance on doubling may be taken from an authoritative names source, such as an Afghan government source or Pashto dictionary; for example, it is usual to see *Hājī* without and *‘Abbās* with the doubled consonant. The doubled *y* consonant is almost always retained, as in *Sayyid* or *Qayyūm*.

- 8 In Afghan names which contain an *izāfah*, it should be romanized as *-e* or *-ye* according to common pronunciation, but generally, *-e* is used if the preceding word ends with a consonant other than silent *heh*, and *-ye* if the preceding word ends with a vowel sound e.g. غر جِصار → *Ghar-e Hiṣār*; قلعة نُو → *Qal'ah-ye Now*. Scholarly sources indicate that *heh* is silent in *darah* and *qal'ah* (thus *darah-ye*, *qal'ah-ye*), but lightly spoken in *kōh* and *chāh* (thus *kōh-e*, *chāh-e*).
- 9 The character sequence خو, where followed by ِ or ى should be romanized *khwā* or *khwī*, although the *w* is either not pronounced, or only weakly so, as in خواجه → *khwājah*.
- 10 Plural nouns ending in *-hā* or *-ān* should always be romanized as a single word, regardless of whether a space appears in a Perso-Arabic script source.
11. Unicode values listed in the tables above are required to ensure standardization and to minimize confusion from competing representations of a given character. It should be noted that the Persian ك (Unicode value 06A9) is recommended rather than the Arabic ك (Unicode value 0643 or FEDA or FED9), the Persian گ (Unicode value 06AF) is recommended rather than گ (Unicode value 06AB) or گ (Unicode value 06B0) or ك (Unicode value 0643 or FEDA or FED9), and the Pashto character خ (Unicode value 0681) is recommended rather than the *heh* with a dot above and a dot below (no Unicode value). For the letter ى in its many variations, care must be exercised to follow this romanization guide's recommendations to eliminate confusion for search engines and software. BGN/PCGN does not use the Unicode encoding FEEF for the character ى in any Afghan word.
- 11 An inventory of letter-diacritic combinations in addition to the unmodified letters of the basic Roman script is:

‘ (U+2018)	’ (U+2019)
Ā (U+0100)	ā (U+0101)
Á (U+00C1)	á (U+00E1)
Ď (U+0044+0031)	ď (U+0064+00031)
Ě (U+0112)	ě (U+0113)
Ê (U+00CA)	ê (U+00EA)
Ħ (U+1E28)	ħ (U+1E29)
Ī (U+012A)	ī (U+012B)
Ñ (U+004E+0304)	ñ (U+004E+0304)
Ō (U+014C)	ō (U+014D)
Ŕ (U+0052+0031)	ř (U+0072+0031)
Ş (U+015E)	ş (U+015F)
Š (U+0053+0304)	š (U+0073+0304)
Ť (U+0054+0031)	ť (U+0074+0031)
Ŧ (U+0162)	ŧ (U+0163)
Ū (U+016A)	ū (U+016B)
Ẑ (U+005A+0327)	ẑ (U+007A+0327)
Ẓ (U+005A+0304)	ẓ (U+007A+0304)
Ẕ (U+005A+0331)	ẕ (U+007A+0331)
ẐH (U+005A+0048+035F)	ẑh (U+007A+0068+035F)

12. The Romanization columns show only lowercase forms but, when romanizing, uppercase and lowercase Roman letters as appropriate should be used.